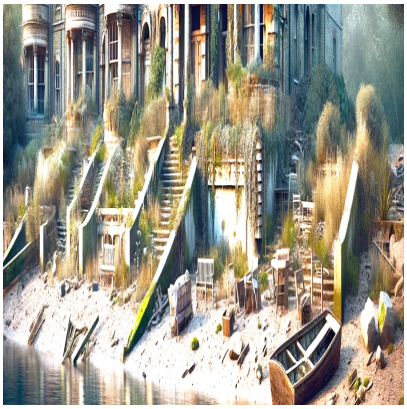

Multi-head attention

Description

My fascination with electronic sound production began when my father brought home a Grundig TS 340 reel-to-reel stereo tape recorder. Like many other recorders it had separate heads for recording and playback, which meant that you could play a recording, add new sounds and feed the combination back through the recording head in real time. I discovered that moody echoes and feedback loops were easy to produce.

The metaphor of tape heads informs the mechanism by which the â??Transformerâ?• method of automated natural language processing (NLP) delivers its convincing performance (as in ChatGPT). Iâ??ve discussed the role of *attention* and NLP attention scores in previous posts. For the mechanical tape head, substitute a set of abstract mathematical NLP functions. The significant Transformer innovation is the use of *multiple attention heads* in processing training data.

Real estate attention heads

Imagine the training data for an NLP model specialising in texts describing property attributes to potential real-estate customers. As described previously, the model will use the semantic encoding data for each token to calculate the relationships between the tokens in a block of text. The model thereby calculates the parts of the text most deserving of attention in a kind of semantic popularity challenge. Hereâ??s a sample of some training text.

> â??Set on a lovely crescent on the shore line of a river, this spacious, four bedroom terraced house, with elegant stone frontage is arranged over two floors and offers spacious accommodation. The property greatly benefits from a lovely open-plan kitchen/dining room with views out over a mature garden to the rear and retains a plethora of period features throughout.â?•

As a potential purchaser, where you direct your attention in this block of text will depend on your interests, and how you priorities those interests. If you are attracted initially to the heritage aspects of a property then the â??phrase elegant stone frontageâ?• may well draw your attention. If you are seeking

somewhere to support a large household then the offer of â??four bedroomsâ?• may draw your attention. If you are in the market for so-called â??affordable housingâ?• then there may be very little in this description that would attract your attention.

To ensure they are applicable to a wide range of language contexts, NLP models need to cater for a range of attention patterns. Multi-head attention attempts to provide for this diversity of interests. One head might focus on the heritage features of the property, while another might focus on size aspects, or its cost. The ChatGPT language model typically has 10 attention heads, catering for 10 different patterns of attention scores across a block of text.

Like a lot of invisible NLP functions the algorithmic identification of features and attention patterns is not as accessible to human interpretation as suggested above. NLP operations involve the mathematical manipulation of the parameters in very large vectors and matrices. Though these systems are tuned to produce outputs legible to human scrutiny, their inner workings are more opaque.

Attention specialists

My description so far suggests that each attention head operates as a specialist. These multiple attention calculations are not prescribed, but the attention pattern in a block of text is influenced by a series of randomly assigned initial conditions for a set of attention matrices. In training, a sliding context window moves across the text sequence, processing a fixed number of tokens at each step. This window determines the current â??focusâ?• of the model, which is crucial for language modelling where the goal is to predict the next word in a sequence.

In models like Transformers, during each step, the attention mechanism computes attention scores (vectors) for all tokens within the context window. These scores determine how much attention each token should receive relative to others when predicting the next token. So the attention vector is effectively a weighting applied to each token, thereby granting it more or less significance as a contributor to the prediction.

Iâ??ve been describing the training phase of an NLP model. The prediction phase, as when you interact with the model conversationally, involves similar attention calculations.

In previous posts I assumed that the final attention pattern in a block of text is a series of simple scores for each token. That made it simpler to represent the attention within a block of text as a heat map. Actually, in these Transformer models the attention patterns are calculated as vectors of the same dimensions as the semantic encoding vector. With multi-head attention the attention parameters for each token are combined to form inputs to the model during training. ChatGPT describes the way attention heads develop their supposed specialisms as â??organic.â?•

Attention mechanisms play a key role in models like ChatGPT. But they are in the company of a larger, complex NLP structure that includes components such as semantic and positional encodings, feed-forward neural networks, layer normalization and fine tuning.

Collaboration

The multi-head attention method has parallels with collaboration within urban settings. Just as different heads in a model might focus on different aspects of the input, individuals or groups in a city might focus on different aspects of a project or problem based on their expertise or interest. These attentions are dynamic and change as members of the group learn from one another and the situation. Iâ??ll discuss this multi-agent analogy in the next post.

Bibliography

- Vaswani, A., N. Shazeer, et al. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA**:** pp. 1-15.

Note

- Featured image is generated by Dall-e from the real estate description I provided in the post. I asked it to depict the setting as post apocalyptic.

Category

1. Artificial Intelligence

Tags

1. attention
2. NLP
3. real estate

Date Created
December 9, 2023
Author
rcoyne99